Regression

Ruwen Qin
Stony Brook University

November 6, 2025

Data used in this module:

- Concrete_Data.csv

Python notebook used in this module

- Regression.ipynb

Considering an entity characterized by an N dimensional feature representation: $\boldsymbol{x} = [x_1, \ldots, x_N]$. We would like to infer measurements of interest, $\boldsymbol{y} = [y_1, \ldots, y_P]$, from the features. Regression is one of the most widely used of all statistical methods to establish the mapping between $\boldsymbol{x}$ and $\boldsymbol{y}$. In a regression model, the feature representation is the input, named predictors or independent variables. The measurements we would like to infer are the outputs and named responses or dependent variables. The responses and predictors can take both numerical and categorical values. This learning module is developed mainly based on references [1, 2].

1 Multiple Linear Regression

Multiple Linear Regression (MLR) handles the special case where there is one response and multiple predictors.

1.1 The Data and Hypothesis Set

Consider a training dataset with M data points. Each data point consists of N predictors and one response. The dataset in the matrix form includes

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_M \end{bmatrix} = \begin{bmatrix} \mathrm{x}_{1,1} & \ldots & \mathrm{x}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathrm{x}_{M,1} & \ldots & \mathrm{x}_{M,N} \end{bmatrix}, \tag{1}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}, \tag{2}$$

The multiple linear regression model relating the response y_i to predictors $x_1, \ldots, x_N$ is

$$y_i = w_0 + w_1 x_{i,1} + \cdots + w_N x_{i,N} + \epsilon_i, \tag{3}$$

for $i = 1, \ldots, M$. ϵ_i in (3) is called the residual, noise, disturbance, or error.

Let

$$\mathbf{Z} = [\mathbf{1}\ \mathbf{X}] = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_M \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{x}_{1,1} & \ldots & \mathrm{x}_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{x}_{M,1} & \ldots & \mathrm{x}_{M,N} \end{bmatrix}, \tag{4}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_M \end{bmatrix}, \tag{5}$$

$$\boldsymbol{w} = \begin{pmatrix} w_0 \\ \vdots \\ w_N \end{pmatrix}. \tag{6}$$

Then, the multiple linear regression model can be expressed as

$$\boldsymbol{y} = \mathbf{Z}\boldsymbol{w} + \boldsymbol{\epsilon}. \tag{7}$$

Assumptions of the linear regression model are:

1. Linearity of the conditional expectation: $\mathrm{E}(y_i|\mathbf{z}_i; \boldsymbol{w}) = \mathbf{z}_i\boldsymbol{w}$;
2. Independent noise: $\epsilon_1, \ldots, \epsilon_M$ are independent;
3. Constant variance: $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$ for all i;
4. Gaussian noise: ϵ_i is normally distributed for all i.

1.1.1 Least Squares Estimator (LSE)

The least-squares estimator (LSE) for $\boldsymbol{w}$ minimizes the square of L_2 distance between the true responses $\boldsymbol{y}$ and their predictions $\widehat{\boldsymbol{y}}$.

The LSE for $\boldsymbol{w}$ is:

$$\widehat{\boldsymbol{w}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{y}. \tag{8}$$

Thereby, the vector of fitted response values is

$$\widehat{\boldsymbol{y}} = \mathbf{Z}\widehat{\boldsymbol{w}}, \tag{9}$$

.

1.1.2 Regression Residuals

Let

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}, \tag{10}$$

and $\widehat{\epsilon}_i$ denote the ith element of $\widehat{\boldsymbol{\epsilon}}$. Then, an unbiased estimator of σ_ϵ^2 is

$$\widehat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{M}\widehat{\epsilon}_i^2}{M-N-1} = \frac{\sum_{i=1}^{M}(y_i-\widehat{y}_i)^2}{M-N-1}. \tag{11}$$

1.1.3 Confidence Interval Estimation of Regression Coefficients

The point estimator of $\boldsymbol{w}$ in (8) is random. An interval estimation of the regression coefficients at the $1-\alpha$ confidence level can be obtained:

$$\widehat{\boldsymbol{w}} \pm t_{1-\alpha/2,M-N-1}\mathrm{se}(\widehat{\boldsymbol{w}}) \tag{12}$$

where $M-N-1$ is the degrees of freedom for the regression residuals, $t_{1-\alpha/2,M-N-1}$ is the t value providing an area of $\alpha/2$ in the right tail of the t-distribution with $M-N-1$ degrees of freedom. If the interval estimate for a coefficient does not contains zero, the corresponding predictor has an impact on the response variable.

1.2 Analysis of Variance

1.2.1 Degree of Freedom

There are degrees of freedom (dof) associated with each of these sources of variation. The dof for regression is N, the number of predictors. The total degrees of freedom is $M-1$. The degrees of freedom for regression residuals is $M-N-1$.

1.2.2 Variance Partitioning

The total variation in $\boldsymbol{y}$ (SST) can be partitioned into two parts: the variation that can be predicted by predictors (SSR) and the variation that cannot be predicted (SSE). Mathematically,

$$\begin{aligned}\mathrm{SST} &= \sum_{i=1}^{M}(y_i-\overline{y})^2 \\ &= \sum_{i=1}^{M}(\widehat{y}_i-\overline{y})^2 + \sum_{i=1}^{M}(y_i-\widehat{y}_i)^2 \\ &= \mathrm{SSR}+\mathrm{SSE}\end{aligned} \tag{13}$$

where $\overline{y}$ is the sample mean computed based on $\boldsymbol{y}$.

1.2.3 Mean Squares of Error

The mean sum of squares is defined to be the ratio of sum of squares to its degrees of freedom.

Therefore, the mean squares total is

$$\text{MST} = \frac{\text{SST}}{M-1}, \tag{14}$$

the mean squares of regression is

$$\text{MSR} = \frac{\text{SSR}}{N}, \tag{15}$$

and the mean squares of error is

$$\text{MSE} = \frac{\text{SSE}}{M-N-1}. \tag{16}$$

which is the unbiased estimator of σ_ϵ^2.

1.2.4 R-Squared, R^2

R-squared, denoted by R^2, is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{17}$$

and it measures the proportion of the total variation in $\boldsymbol{y}$ that can be linearly predicted by predictors $\mathbf{X}$.

1.2.5 F Test

F test of goodness-of-fit uses the test statistic F:

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{18}$$

to examine the hypotheses:
H_0: $w_0 = \cdots = w_N = 0\,; \forall j$
H_1: $\exists\, w_j \neq 0\ j \in \{0, \ldots, N\}$

Rejecting the null hypothesis means a regression relationship exists between the predictors and the response variable. The p value corresponding to this test statistic is the probability that a random value drawn from the F distribution with N and $M-N-1$ degrees of freedom is greater than the test statistic. The larger the test stastistic, the smaller the p value and so the chance of making a mistake in rejecting the null hypothesis.

1.3 Prediction Accuracy

The key concept associated with measuring forecast accuracy is forecast error. Various metrics are used to determine how well a particular forecasting method performs. Given $\boldsymbol{x}_i$ from the test set, a regression model predicts the response $\widehat{y}_i$. Multiple metrices are defined for assessing the prediction error for a numerical response:

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{\sum_{i=1}^{S} |y_i - \widehat{y}_i|}{S}, \tag{19}$$

Mean Square Error (MSE)

$$\text{MSE} = \frac{\sum_{i=1}^{S} (y_i - \widehat{y}_i)^2}{S}, \tag{20}$$

Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{S} (y_i - \widehat{y}_i)^2}{S}}, \tag{21}$$

where S is the size of test dataset.

2 An Example of Multiple Linear Regression

2.1 The Problem

Concrete has 8 attributes:

1. Cement – quantitative – kg in a m^3 mixture
2. Blast Furnace Slag – quantitative – kg in a m^3 mixture
3. Fly Ash – quantitative – kg in a m^3 mixture
4. Water – quantitative – kg in a m^3 mixture
5. Superplasticizer – quantitative – kg in a m^3 mixture
6. Coarse Aggregate – quantitative – kg in a m^3 mixture
7. Fine Aggregate – quantitative – kg in a m^3 mixture
8. Age – quantitative – Day (1∼365)

We would like to build a regression model that either uses these hand-crafted features or extracted ones to predict concrete's compressive strength in megapascal (MPa).

2.2 Ordinary Linear Regression

2.2.1 The Attained Model

We fit a linear regression model that predicts the compressive strength of concrete using the 8 principal components extracted from the eight concrete attributes as predictors. The 95% confidence intervals of the coefficients for C1 and C2 contain zero, indicating these two Cs may not be important predictors.

```
      Coefficient  Margin of error
w0        35.666        0.722
w1        -0.156        0.474
w2        -0.468        0.605
w3         7.375        0.636
w4         1.329        0.719
w5        -5.037        0.739
w6         7.749        0.809
w7        -7.370        1.680
w8        11.752        4.150
```

2.2.2 Residuals

First, we visually checked the residuals from the regression to make sure assumptions about residuals for the regression model hold. If needed, quantitative test results for the regression model can be calculated and they are more precise information.

Figure 1 is the plots of residuals attained on the training and testing datasets. The plots indicate residuals are independent.

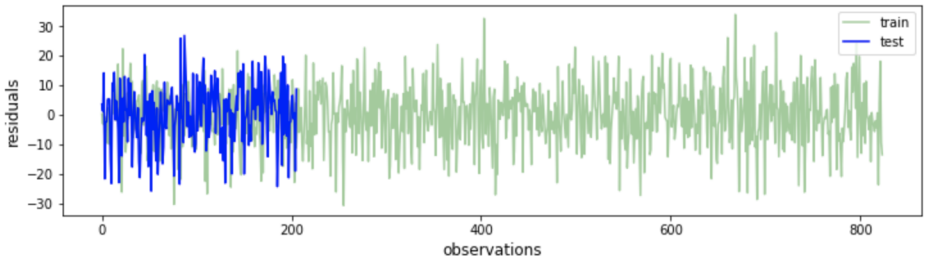

Figure 1: Residuals

Figure 2 shows the distributions of residuals obtained on the training and the testing datasets. The distributions are close to a normal distribution.

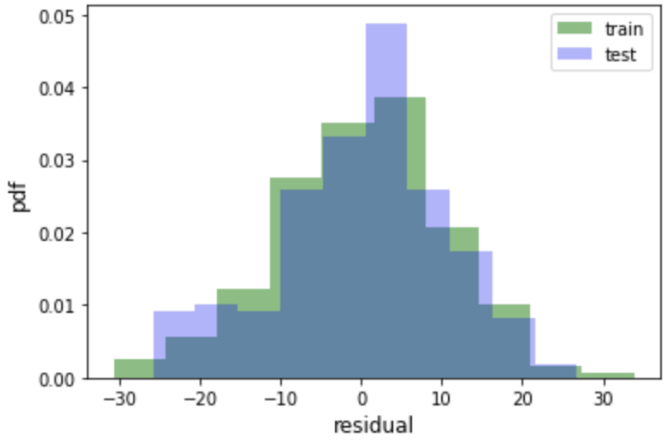

Figure 2: Residual Distribution

Figure 3 shows residuals at predicted response values. The plot shows the distribution of residuals increases when the predicted response value increases. Residuals are not homogeneously distributed. This raises an alarm. Therefore, the current linear model might not be the best choice. We will revisit this issue in later sections. In the following, we continued evaluating this linear model.

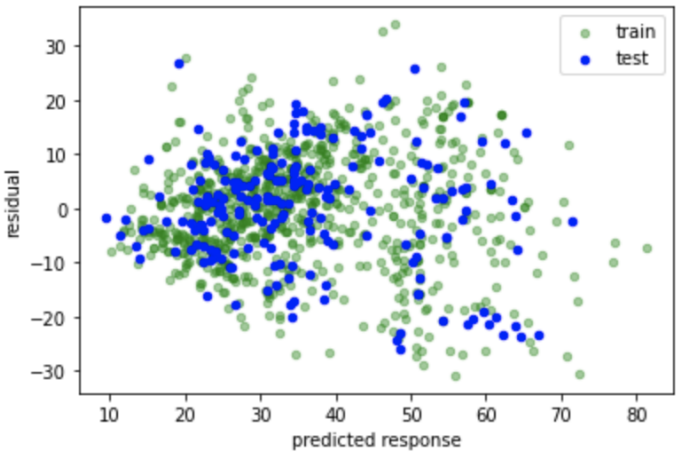

Figure 3: Residuals vs. predicted response

2.2.3 Goodness-of-Fit

In this example, the average R^2 achieved on the training dataset using a five-fold cross validation is 56.41%, meaning the amount of data variation captured by the model. R^2_{adj} is 55.97%, slightly smaller than R^2, indicating the model is not that complex.

```
AIC:6,051
BIC:6,093
The average R squared value for training: 56.41%
The average adjusted R squared value for training: 55.97%
```

In testing the model on a test dataset, the accuracy is measured as below:

```
MAE: 7.618
MSE: 99.141
```

```
RMSE: 9.957
R Square: 0.616
```

3 Polynomial Linear Regression

Polynomial linear regression fits a polynomial function of predictors $\boldsymbol{x} = [x_1, \ldots, x_N]$ to predict the response y.

For example, a polynomial in three predictors $[x_1, x_2, x_3]$ with degree of 2 has 10 ($=C_2^{3+2}$) polynomial terms: 1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2.

We can fit a polynomial regression function using the linear regression model.

4 Binary Logistic Regression

4.1 The Data and Hypothesis Set

A training dataset contains M data points. Each data point consists of the observations of N predictors and a binary response. Binary logistic regression models the conditional probability that a binary response is 1 given predictors and their coefficients. The binary logistic regression can take the following form:

$$p_i = p(y_i = 1|\mathbf{z}_i; \boldsymbol{w}) = \text{sigmoid}(\mathbf{z}_i\boldsymbol{w}) = \frac{1}{1 + \exp(-\mathbf{z}_i\boldsymbol{w})}, \tag{22}$$

where $\boldsymbol{w}$ is the vector of regression coefficients

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_N \end{bmatrix}, \tag{23}$$

and $\mathbf{z}_i = [1\ \boldsymbol{x}_i]$. We thereby define $\mathbf{Z}$ by appending a column of ones to the left of $\mathbf{X}$:

$$\mathbf{Z} = [\mathbf{1}\ \mathbf{X}] = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_M \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{x}_{1,1} & \ldots & \mathrm{x}_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{x}_{M,1} & \ldots & \mathrm{x}_{M,N} \end{bmatrix}. \tag{24}$$

By learning from the training dataset, the regression coefficients $\boldsymbol{w}$ can be estimated, for example, using maximum likelihood estimation.

4.2 An Example

Logistic regression is commonly used in classification problems. Let's consider a problem similar to the concrete example we studied. The response y is a binary variable indicates the class of a concrete slab: "defective" vs. "non-defective". y takes the value 1 meaning defective, and 0 meaning non-defective. We split the data by 80-20. That is, 80% of the data are used for training a model and 20% for testing the developed model.

The maximum likelihood estimates of regression coefficients are:

```
  Coefficient
w0  -3.253
w1  -1.644
w2   0.924
w3  -1.609
w4   2.284
w5   2.062
w6  -5.268
w7   1.082
w8  -2.021
```

The predicted class is

$$\widehat{y}_i = \begin{cases} 1 & \text{if } p_i > 0.5 \\ 0 & \text{if } p_i \leq 0.5. \end{cases} \tag{25}$$

Figure 4 is the confusion matrix. Among the 136 true zeros, 123 are predicted correctly, and 13 are predicted mistakenly. Among the 65 true ones, 57 are predicted correctly, and 8 are predictably mistakenly. The accuracy is 89.55%(= $(123 + 57)/(123 + 13 + 9 + 57)$).

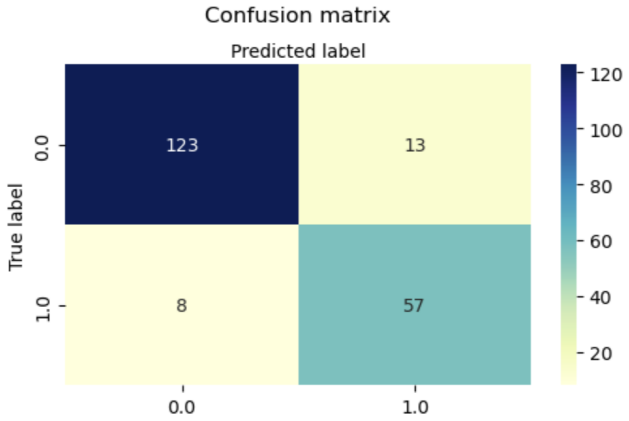

Figure 4: Confusion matrix

5 Closing Remark

In this learning module, we primarily introduced several popularly used regression models falling in the category of *Generalized Linear Models* (GLM). A GLM maps a linear combination of predictors $\boldsymbol{x}_i\boldsymbol{w}$ to the expected value of response $\mathrm{E}[y_i|\boldsymbol{x}_i]$ through a link function $l(\boldsymbol{x}_i\boldsymbol{w}) = \mathrm{E}[y_i|\boldsymbol{x}; \boldsymbol{w}]$.

We considered various distributions of the response in regression: the response in the multiple linear regression follows a univariate Gaussian distribution, and the Bernoulli distribution in the binary logistic regression. There are a few more distributions of responses that this learning module did not cover, such as responses following binomial distribution or Poisson distribution. Regression models for those can be developed similarly.

References

[1] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction.* MIT Press, 2022.

[2] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics.* MIT Press, 2023.